Lieff
Cabraser
Heimann&
Bernstein
Attorneys at Law

SUSMAN GODFREY L.L.P.
A REGISTERED LIMITED LIABILITY PARTNERSHIP

CDAS

January 13, 2025

VIA ECF

Hon. Ona T. Wang
Daniel Patrick Moynihan
United States Courthouse
500 Pearl St.
New York, NY 10007

RE: *Authors Guild v. OpenAI Inc.*, (No. 1:23-cv-08292-SHS)
Alter v. OpenAI Inc., (No. 1:23-cv-10211-SHS)

Dear Judge Wang:

Pursuant to Rule II(b) of Your Honor's Individual Practices, Plaintiffs seek a conference regarding OpenAI's failure to produce source code and other information related to OpenAI's downloading, use, and modification of the text repositories that train the large language models ("LLMs") powering ChatGPT. OpenAI objects to providing the requested source code and information as not responsive to Plaintiffs' Requests for Production ("RFPs"). OpenAI is wrong and must produce the requested source code and related information for three reasons: (1) because the source code and related information are directly responsive to many of Plaintiffs' RFPs, (2) because the source code and related information are relevant, and (3) because the source code and related information are proportional to the needs of the case.

Background. On August 15, OpenAI said it would provide source code that was at a minimum responsive to Plaintiffs' RFP No. 79. Ex. 1 at 14. It took OpenAI more than three months to make its source code available for inspection. OpenAI finally provided some source code in late November, and Plaintiffs inspected the source code in December. Shortly after Plaintiffs' source code inspection, Plaintiffs promptly identified numerous omissions in the source code and asked OpenAI to provide additional source code and related information, including:

A. Source code related to books datasets known as LibGen, Books1, Books2, and other datasets that Plaintiffs have a good faith basis to believe contain books, including the code that downloaded these datasets, discovered the lists of books to download, scraped the webpages, and filtered and processed the data.

B. Source code related to GPTBot, including how it makes web requests, discovers pages to crawl, and processes the downloaded data.

C. Model architecture code for GPT-3, GPT-4, and GPT-4o that clearly defines all training layers and functions.

D. Model training code that iteratively optimizes the model parameters, including code that feeds training data into the model and optimizes functions used to train the model.

E. Model hosting and inference code, including code that processes user prompts, calls the model, and returns a response.

F. Content moderation code, including code that detects whether a response should be returned to a ChatGPT user.

G. Scripts that include the terms "book," "books," "copyright," or "license".

H. Training datasets present and referenced in the source code that have not been produced.

I. Documents referenced in the source code that relate to the source and copyrightability of training data.

Despite several emails from Plaintiffs, OpenAI failed to confirm that it would produce the requested source code and related information. On the parties' January 9 meet and confer, OpenAI objected to the responsiveness of the source code and related information. Plaintiffs disagreed, cited 14 RFPs to which the requested information is responsive, and asked OpenAI to confirm that it would provide the requested information. On January 10, in an abundance of caution, Plaintiffs served additional RFPs seeking the source code. Because OpenAI has not confirmed that it will provide the additional source code and related information, the parties are at impasse.

The source code and related information are responsive. OpenAI objects to providing the requested source code because this information is allegedly not responsive to Plaintiffs' RFPs. OpenAI's argument is unfounded. As Plaintiffs explained to OpenAI on the parties' meet and confer, there are at least 14 RFPs that cover the requested source code and information, including:

- RFP No. 17, which seeks documents[1] sufficient to determine any manner that OpenAI's LLMs have referenced or accessed works of fiction and non-fiction. Ex. 2 at 7. At least source code requests A, G, and H are responsive to RFP No. 17.
- RFP No. 18, which seeks documents sufficient to identify the electronically stored information through which OpenAI accessed commercial works of fiction and non-fiction to train ChatGPT. *Id.* At least source code requests A, G, and H are responsive to RFP No. 18.
- RFP No. 31, which seeks documents sufficient to determine the source material for Books1 and/or Books2. *Id.* at 9. At least source code request A is responsive to RFP No. 31.
- RFP No. 39, which seeks documents related to OpenAI's curation of data and the types of data OpenAI uses to train its LLMs, including any differences in the ways that OpenAI's LLMs ingest, process, or output data and the decision to include or exclude certain training data. Ex. 3 at 8. At least source code requests A, C, D, E, G, H, and I

[1] As defined in Plaintiffs' RFPs, "documents" means "all materials within the scope of Federal Rule of Civil Procedure 34," which includes electronically stored information, data, and data compilations. Ex. ___; *see also* Fed. R. Civ. P. 34(a)(1). Rule 34 is broad and encompasses source code. *See, e.g.*, Columbia *Pictures, Inc. v. Bunnell*, 245 F.R.D. 443, 447 (C.D. Cal. 2007) ("Rule 34(a)(1) is intended to be broad enough to cover all current types of computer-based information, and flexible enough to encompass future changes and developments.") (citation omitted).

are responsive to RFP No. 39.

- RFP No. 41, which seeks documents related to OpenAI's modification of any parameters for tuning or limiting OpenAI's LLMs' outputs to avoid copyright infringement. *Id.* At least source code requests D, E, F, and I are responsive to RFP No. 41.
- RFP No. 55, which seeks documents related to OpenAI's use of text repositories including LibraryGenesis, CommonCrawl, and Internet Archives to train OpenAI's LLMs. *Id.* at 11. At least source code requests A, B, C, G, and H are responsive to RFP No. 55.
- RFP No. 60, which seeks documents describing the way that OpenAI's LLMs use training data. *Id.* at 12. At least requests C and G are responsive to RFP No. 60.
- RFP No. 61, which seeks documents sufficient to show all audiobooks OpenAI transcribed. *Id.* At least source code requests A is responsive to RFP No. 61.
- RFP No. 65, which seeks documents related to OpenAI's use of web crawler permissions in an effort to comply with copyright law. *Id.* At least source code request B is responsive to RFP No. 65.

The source code and related information are relevant. Plaintiffs allege that OpenAI downloaded repositories of pirated text and knowingly used these repositories to train ChatGPT and the LLMs that underly ChatGPT. *See, e.g.*, First Am. Compl. ¶¶ 115-116, 169. Source code and information showing which datasets OpenAI downloaded, how OpenAI scraped the Internet for these datasets, how OpenAI downloaded these datasets, how OpenAI used these datasets during training, how OpenAI finetuned its models, and how OpenAI modified (or did not modify) its models to avoid outputting copyrighted information is all highly relevant to Plaintiffs' claims.

The source code and related information are proportional to the needs of the case. The requested source code is proportional because the benefits of the source code to Plaintiffs outweigh any purported burden to OpenAI. *See* Fed. R. Civ. P. 26(b). Plaintiffs need additional source code and information to identify the books datasets that OpenAI uses in training and to understand how OpenAI uses these datasets, and neither the training data nor the source code OpenAI has provided thus far provide this critical information. OpenAI has not explained why the requested source code is not proportional to the needs of the case, has made no demonstration regarding the burden of compiling and producing it, and, in any event, the burden of providing this information is minimal because OpenAI is likely to provide much of this source code in its other cases.

Because the requested source code is responsive, relevant, and proportional to the needs of the case, OpenAI must provide the requested source code and related information described above.

Sincerely,

LIEFF CABRASER HEIMANN & BERNSTEIN LLP	SUSMAN GODFREY LLP	COWAN, DEBAETS, ABRAHAMS & SHEPPARD LLP
/s/ Rachel Geman Rachel Geman	/s/ Rohit Nath Rohit Nath	/s/ Scott J. Sholder Scott J. Sholder